3-3

УДК 004.8

А.С. Звенигородский, А.И. Шевченко
Институт информатики и искусственного интеллекта
ГВУЗ «Донецкий национальный технический университет», г. Донецк, Украина
Украина, 83050, г. Донецк, ул. Б. Хмельницкого, 84, г. Донецк, zas@suiai.edu.ua

Трансляция естественно-языковых запросов к базе данных в SQL-запросы

A.S. Zvenigorodsky, A.I. Shevchenko
Institute of Informatics and Artificial Intelligence
of Donetsk National Technical University, Donetsk, Ukraine
Ukraine, 83050, c. Donetsk, B. Khmelnitskiy st., 84, zas@suiai.edu.ua

The Transformation of Natural Language Database Gueries to SQ-Gueries

О.С. Звенігородський, А.І. Шевченко
Інститут інформатики і штучного інтелекту
ДВНЗ «Донецький національний технічний університет», м. Донецьк, Україна
Україна, 83050, м. Донецьк, пр. Б. Хмельницького, 84, zas@suiai.edu.ua

Трансляція природно-мовних запитів до бази даних у SQL-запити

В статье рассматриваются проблемы трансляции естественно-языковых запросов в запросы SQL-языка. Предлагаются методы и алгоритмы, позволяющие интерпретировать ЕЯ-запрос ограниченным набором шаблонов SQL-запроса.
Ключевые слова: естественно-языковой запрос, SQL-запрос, база данных, смысл текста.

The problems of transformation of natural language database queries to SQL queries are discussed in the article. Transformation methods and algorithms for constrained set of SQL pattern are proposed.
Key Words: natural language queries, SQL queries, database, text meaning.

У статті розглядаються проблеми трансляції природно-мовних запитів у запити SQL-мови. Пропонуються методи та алгоритми, що дозволяють інтерпретувати ПМ-запит обмеженим набором шаблонів SQL-запиту.
Ключові слова: природно-мовний запит, SQL-запит, база даних, зміст тексту.

Введение

У многих информационных систем основным ядром является SQL-ориентированные базы данных (БД). Для обычного пользователя, не разработчика БД, знание языка SQL и всех тонкостей строения базы данных требует дополнительных усилий и не всегда оправдано. Обычно для пользователя БД, если он хочет получить интересующую его информацию, предлагаются некоторые формы, которые надо заполнить. По введенным данным программа формирует SQL-запрос. Изначально эти формы создаются для определенного запроса, например, получения данных о сотрудниках (сотруднике). Чтобы не выводить все данные, пользователь заполняет определенные поля формы и затем получает информацию. Это, с одной стороны, хорошо, система запрашивает только релевантные данные. С другой – плохо, потому, что пользователю требуется время, чтобы найти и открыть нужную форму, затем не ошибиться в задании данных (подспорьем является выпадающие списки с правильными значениями,

но если список большой, то понадобится время, чтобы найти требуемую строку и кликнуть на ней). При ЕЯ-запросе необходимость в формах отпадает, пользователь указывает только ту информацию, которая его интересует, чем экономит свое время. Поэтому задача построения естественно-языкового (ЕЯ) пользовательского интерфейса, позволяющего взаимодействовать с БД на естественном языке, является актуальной. **Целью данной статьи** является уменьшение неоднозначности смысловой интерпретации естественно-языковых запросов к БД.

Состояние проблемы

Наиболее распространенными подходами создания пользовательского интерфейса трансляции ЕЯ-запросов в запросы SQL являются подходы, основанные на синтаксическом, семантическом анализах и шаблонах [1]. Синтаксическое представление запроса строится на анализе членов предложения: подлежащего, сказуемого, прямого дополнения и т.п., которые определяются с помощью морфологических характеристик. Но для однозначной интерпретации этих характеристик недостаточно, так как они не отражают смысла запроса в полной мере.

Второй подход, основанный на семантике, ближе к смыслу запроса [1]. В нем используется синтаксическая информация из предыдущего подхода, а также информация из тезаурусов. С помощью тезаурусов строится семантическое представление запроса. Основная задача при этом – отсечь ненужные смыслы, постараться выделить с помощью синтаксических связей достоверные семантические конструкции. Основная сложность данного анализа – типичные естественно-языковые запросы, которые, как правило, не имеют правильных синтаксических конструкций, что обусловлено вольным словоизменением, большим количеством имен собственных и сокращений, игнорирования правил пунктуации [1].

Третий подход к анализу естественно-языковых запросов основан на шаблонах. Представителем шаблонного подхода является система English Query от Microsoft, основанная на синтаксически-ориентированных шаблонах, связываемых с моделью предметной области, и через нее – со схемой базы данных. Система имеет ряд существенных недостатков:

1) пользователь может употреблять только правильные входные предложения;

2) в процессе эксплуатации неизбежно приходиться модифицировать и расширять знания о языке и окружающем мире, заложенные в ЕЯ-интерфейсе при его разработке, что является трудоёмкой задачей, которая не может быть выполнена пользователями системы;

3) если система не может найти ответ на запрос, то она не способна объяснить пользователю причину неудачи;

4) позволяет строить естественно-языковые интерфейсы только для английского языка и работает только с Microsoft SQL Server.

Среди отечественных технологий, позволяющих строить ЕЯ-интерфейсы к реляционным базам данных, наиболее известной является технология InBase [2], [3], основанная на шаблонном подходе. Необходимо заметить, что на сегодняшний день не существует системы, адекватно реализующей шаблонный подход к русскоязычным ЕЯ-запросам. Таким образом, рассмотренные методы не решают задачу понимания русскоязычных запросов к БД и необходим поиск новых подходов.

Исходные положения

Определим отличительные черты процесса понимания ЕЯ-запросов. Во-первых, ЕЯ-запрос формируется в диалоге пользователя с БД, поэтому некоторые се-

мантические составляющие непосредственно не содержатся в запросе, а подразумеваются или устанавливаются по результатам предыдущего диалога. Во-вторых, должна ли система задавать наводящие вопросы в случае грамматических и синтаксических ошибок. Это увеличивает время на получение заданной информации и раздражает пользователя тем, что надо отвечать на лишние вопросы. Поэтому мы приняли следующие начальные условия:

– в ЕЯ-запросе допускается произвольный порядок слов;

– в ЕЯ-запросе допускаются грамматические ошибки, не искажающие смысл запроса;

– в ЕЯ-запросе допускаются длинные и короткие выражения, например, «показать адреса всех сотрудников»; «все адреса» или просто «адреса»;

– система не задает наводящих вопросов, если не удается построить SQL-запрос, то система отвечает, что не понимает и советует переформулировать запрос или проверить грамматику;

– если пользователь получит не ту информацию, которую ожидал, то это ошибка пользователя, а не системы.

Постановка задачи

Идея. Объект имеет смысл, если мы можем мыслить о нем или его свойствах путем формулирования вопросов (запросов) на ЕЯ.

В [4] предложена модель, согласно которой смысл имеет проблемная область ситуаций (ПОС). Одной из оставляющих ПОС являются образы предметной области. Существует также языковая проблемная область ситуаций (ЯПОС) – формальная система для выражения ситуаций и образов предметной области средствами ЕЯ [4]. Суть ЯПОС в том, что образ и его свойства не зависят от естественного языка. В естественном языке они обозначаются словами, словосочетаниями, порядок и построение которых определяется правилами грамматики, но, как известно, грамматически правильное предложение может быть бессмысленным. Поэтому мы считаем, что текст или ЕЯ-выражение смысла не имеют. Они содержат языковые конструкции, которые указывают на смысловые составляющие определенной предметной области, которые существуют независимо от языка. Таким образом, возможны два варианта анализа запросов на ЕЯ:

ЕЯ-ЗАПРОС – СМЫСЛЫ БД – SQL-ЗАПРОС,
SQL-ЗАПРОС – СМЫСЛЫ БД – ЕЯ-ЗАПРОС.

В первом случае предполагается, что нескольким ЕЯ-запросам соответствует один SQL-запрос, во втором – одному SQL-запросу соответствует несколько ЕЯ-запросов, т.е. процесс анализа можно проводить в прямом направлении и в обратном, или одновременно в обоих. Отсюда следует, что необходимо решить следующий ряд задач:

1. Определить смысловые составляющие БД.
2. Построить модель интерпретации SQL-запроса смысловыми составляющими БД.
3. Построить модель интерпретации ЕЯ-запроса смысловыми составляющими БД.

Естественно-языковая модель запроса к базе данных

Смысловые составляющие рассмотрим на примере базы данных отдела кадров табл. 1.

Пусть есть некоторое множество часто задаваемых запросов относительно адреса. Этими запросами могут быть: «Где живет Иванов?», «Где живут работники?»,

«Список адресов работников», «Все адреса работников», «Адреса всех работников бухгалтерии», «Адреса работников», «Адреса всех программистов», «Адреса Ивановых?».

Таблица 1 – Таблица «Staff» базы данных отдела кадров

№	Surname	Address	Appointment	Phone	Salary
1	*Иванов*	*Донецк, ул. Кирова 17, кв. 45*	*Программист*	*4564715*	*1500*
2	*Рыбкин*	*Донецк, ул. Горького, 5*	*Бухгалтер*	*4568774*	*2000*

Все эти вопросы покрываются тремя шаблонами SQL-запросов.

1. Запрос на всю информацию.
   SELECT, Surname, Address, Appointment, Phone, Salary
   FROM Staff
2. Запрос на все данные одного атрибута (столбца).
   SELECT Address
   FROM Staff
3. Запрос на данные одного столбца с простым условием.
   SELECT Address
   FROM Staff
   WHERE Surname = Иванов

Ограничимся этими запросами. По определению SQL-запрос не может иметь несколько значений, т.е. его синтаксис совпадает с семантикой. В указанных выше SQL-запросах сущностями БД является таблица с данными, столбцы таблицы и значения свойств (данные в ячейках столбов). В самих запросах они обозначаются как имя таблицы, имена столбцов и данные определенного типа. Отметим, что изначально SQL-запрос не существует, независимо от программиста, программист должен его создать. Смоделируем процесс создания программистом SQL-запроса по следующему заданию: «Составьте список адресов всех Ивановых». В данном случае программист должен определить шаблон SQL-запроса и все его параметры.

Основным параметром в SQL-запросе является параметр *FROM*. Его значением является имя таблицы *Staff*. На это имя программиста наталкивает слово «адресов» или словосочетание «список адресов», так как он помнит, что в БД отдела кадров есть столбец, в котором указываются адреса сотрудников. Также на это имя его может натолкнуть осознание того, что он в данный момент работает с БД отдела кадров. Другими словами, это имя извлекается из головы программиста или контекста на основании мышления, так как прямого указания на него в запросе нет. На то, что это шаблон № 3, указывает слово «Ивановых», так как в сознании программиста оно связано со значением *Иванов* и знаний о том, что в SQL-запросе можно указывать условия для сокращения выборки. Это условие указывается в операторе *WHERE* по определенным правилам. В данной статье рассматривается только правило на равенство. Для параметра *SELECT* в сознании программиста на столбец с именем *Address* указывает слово «адресов». Для употребления параметра *WHERE* и условием на равенство со значением *Surname = Иванов* указывает слово «Ивановых», так как значение *Иванов* может быть только в этом столбце и не может быть в других, например, *Post, Phone, Salary*. Можем сказать, что *Surname* и *Иванов*, тоже находится в контексте. Во всех рассмотренных случаях нет прямого указания на параметры SQL-запроса. Все они извлекаются программистом из контекста. Таким образом, в SQL-запросе явно указываются сущности БД, а в ЕЯ-запросе находятся косвенные указатели на них. Следовательно, в контексте системы должны быть знания, как перейти от слов и выражений ЕЯ к значениям параметров SQL-запроса и знания о языке SQL.

Уточним теперь сущности БД в системе трансляции одних запросов в другие. Воспользуемся определениями из [5] и распространим их на БД.

Образ – строка или запись в таблице БД.

Класс образов – таблица в БД.

Атрибут – столбец таблицы (класса образов).

Значение атрибута – код, сформированный по заданным правилам и хранящийся в ячейке столбца (последовательность ASCII символов, число и другие типы данных).

Естественно-языковая форма (ЕЯ-форма) – слова и последовательности слов, употребляемые в предметной области для обозначения смысловых составляющих базы данных.

Контекст – знания о способах определения кода таблицы, кодов столбцов и значений атрибутов образа, с которыми связаны ЕЯ-формы, протокол диалога и негласные соглашения, основанные на принципах функционирования реляционных БД.

Для того чтобы «мыслить» над введенными сущностями БД, примем, что в системе они обозначаются уникальными кодами, не совпадающими с ЕЯ-составляющими и именами в SQL-запросах.

Обобщая вышеизложенное, ЕЯ-модель запроса к БД представим как совокупность множеств ЕЯ-форм, которые ассоциируются с образами, атрибутами и значениями атрибутов БД на основе правил и данных контекста. Для рассматриваемых SQL-шаблонов модель представляется семеркой:

$$NL = (NLFW, W, NLFY, Y, S, Q, C),$$

где *NLFW* – множество основ ЕЯ-форм и морфологические правила для значений атрибутов; *W* – множество правил определения соответствий ЕЯ-форм значениям атрибутов БД; *NLFY* – множество основ ЕЯ-форм и морфологические правила для атрибутов; *Y* – множество правил определения соответствия ЕЯ-форм атрибутам БД; *S* – синтаксические правила для всего ЕЯ-запроса для указания пользователю на ошибки; *Q* – знания о языке SQL и его версиях; C – контекста (знания о БД, правила ведения диалога и предыдущие результаты диалога).

Множество *NLFW* состоит из подмножеств, в которых указываются основы слов и морфологические признаки, например, для слов «Иванов», «Иванова», «Ивановых», которые могут быть в ЕЯ-запросе.

Правила множества *W* определяют, что в БД словам «Иванов», «Иванова», «Ивановых» соответствует значение *Иванов*, которое может быть в одной или нескольких ячеек в столбце, где хранятся фамилии. Для SQL-запроса БД отдела кадров этот столбец имеет имя *Surname*.

Множество *NLFY* состоит из подмножеств, в которых указываются основы слов и выражения (сочетания слов) и морфологические признаки, например, для слов «где живет», «где живут работники», «список адресов работников», «все адреса работников», «адреса работников», которые могут быть в ЕЯ-запросе.

Правила множества *Y* определяют, что в БД словам «где живет», «где живут работники», «список адресов работников», «все адреса работников», «адреса работников» соответствует столбец, в котором хранятся адреса. Для SQL-запроса БД отдела кадров этот столбец имеет имя *Address*.

В целом процесс формирования SQL-запроса по ЕЯ-запросу представлен на рис. 1.

В дальнейшем будут разрабатываться модели на запросы к нескольким столбцам и таблицам и с более сложными условиями.

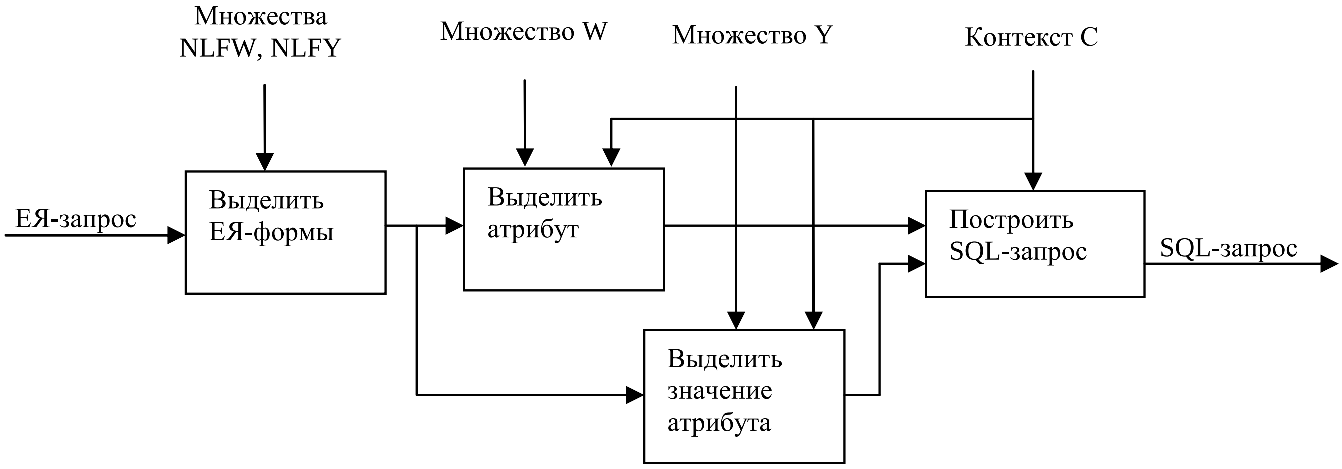


Рисунок 1 – Процесс трансляции ЕЯ-запроса в SQL-запрос

Выводы

1. В ЕЯ-запросе находятся косвенные указатели на параметры SQL-запроса, которые интерпретируются только через контекст.

2. Предложенный подход не накладывает ограничений на содержание и порядок слов ЕЯ-запроса.

3. Подмножества модели разрабатываются отдельно, что позволяет наращивать эффективность системы в процессе эксплуатации и использовать их в разных БД.

4. Трудности реализации данного подхода связаны в основном с выделением имен собственных: фамилий, географических названий и т.п., количество которых может быть значительным.

Литература

1. Найханова Л.В. Методы и алгоритмы трансляции естественно-языковых запросов к базе данных в SQL-запросы : [монография] / Найханова Л.В., Евдокимова И.С. – Улан-Удэ : Изд-во ВСГТУ, 2004. – 148 с.: ил.
2. Жигалов В.А. InBASE: технология построения ЕЯ-интерфейсов к базам данных / Жигалов В.А., Соколова // Труды Международного семинара Диалог’2001 по компьютерной лингвистике. – Аксаково, 2000. – С. 123-135.
3. Хорошевский В.Ф. Обработка естественно-языковых текстов: от моделей понимания к технологиям извлечения знаний / В.Ф. Хорошевский // Новости искусственного интеллекта. – 2002. – № 6. – С. 19-26.
4. Звенигородский А.С. Кибернетические основы процесса понимания смысла текста / А.С. Звенигородский // Искусственный интеллект. – 2010. – № 4. – С. 82-89.
5. Святогор Л. Определение понятия «смысл» через онтологию / Л. Святогор, В. Гладун // Семантический анализ текстов естественного языка : XVth International Conference «Knowledge-Dialogue-Solution» KDS 2009. – Varna, Bulgaria, June-July 2009.

Literatura

1. Naihanova L.V. Metody i algoritmy translyacii estestvenno-yazykovyh zaprosov k baze dannyh v SQL-zaprosy: Monografiya. Ulan-Ude: Izd-vo VSGTU. 2004. 148 s.
2. Zhigalov V.A. Trudy Mezhdunarodnogo seminara Dialog’2001 po komp’juternoi lingvistike. Aksakovo. 2000. S. 123-135.
3. Horoshevskii V. F. Novosti iskusstvennogo intellekta. 2002. № 6. S. 19-26.
4. Svyatogor L. Semanticheskii analiz tekstov estestvennogo yazyka. XVth International Conference “Knowledge-Dialogue-Solution” KDS 2009. Varna. Bulgaria. June-July 2009.
5. Zvenigorodskii A.S. Iskusstvennyi intellekt. 2010. № 4. S. 82-89.

Статья поступила в редакцию 06.06.2012.